# Automatic Generation of Corpus-Based Exercises Using Generative AI

**Adrian Jan Zasina**
Institute of Czech and Deaf Studies
Charles University
Prague, Czech Republic
adrian.zasina@ff.cuni.cz

## Abstract

This study explores the automatic generation of corpus-based language exercises using a generative AI model Corpus Linguist. It focuses on the interaction between the language model and corpus data, detailing a workflow in which collocation and translation patterns are extracted from a tagged corpus and structured prompts are constructed to guide the model in producing sentence-level exercises. The generated exercises reveal both the potential and the current limitations of AI-driven approaches. Challenges include inconsistency in corpus data use, and choosing appropriate translation equivalents. These observations highlight the necessity of careful design and critical evaluation when integrating generative models with corpus-based language materials. By analysing these processes from a computational linguistics perspective, this study contributes to understanding how generative AI can interact with structured linguistic data, informing future applications in automated language resources.

Keywords: AI, corpus, corpus-based exercises, DDL

## 1 Introduction

Since November 2022, when ChatGPT from OpenAI was released, new language models using generative artificial intelligence (GenAI) have appeared. These are no longer simple chatbots but more advanced tools that allow users not only to engage in conversation but also to create images and videos, and perform data analysis. Thanks to this development, it has become possible, for the very first time, to link corpus data with GenAI to create human-like queries for corpora, instead of relying on traditional queries in Corpus Query Language (CQL), which require professional knowledge of CQL syntax. Combining corpus methods with GenAI opens new possibilities in language analysis, enabling us to evaluate how GPT models interact with corpus data.

One reason for this synergy is the integration of corpus data and GenAI with data-driven learning (DDL), which relies on authentic data in language learning (Johns, 1991). In the early 1990s, DDL faced challenges due to the lack of user-friendly tools that could attract learners. Some studies (Vyatkina, 2020, pp. 362–363; Zasina, 2022, p. 126) highlight that learners benefit from corpus activities, however, the tools available at the time were often too complex. The advent of ChatGPT has enabled fast communication and brought substantial computational capacities. Finally, the creation of Corpus Linguist GPT model (Milička & Machálek, 2024) in 2024 has made it feasible to use corpus data in a user-friendly environment. Some corpus linguists are aware of the limitations of existing corpus interfaces, therefore, new studies (Cheung & Crosthwaite, 2025) combining these two sources are beginning to appear.

However, there is little evidence on the interaction between language models and corpus data in terms of DDL practice (cf. Zasina, 2025). Therefore, this study attempts to explore the automatic generation of corpus-based exercises for learners using a GenAI model. First, attention is paid to prompt crafting along with the co-star framework (Lin, 2025), which enables better results. Second, the study focuses on the interaction between the Corpus Linguist model and corpus data. The aim is to evaluate both the potential and current limitations of combining GenAI and corpus data in

*Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)*, pages 80–86
November 20-22, 2025 ©2025 Association for Computational Linguistics

terms of corpus searching and the generation of corpus-based exercises.

## 2 Data

For the purpose of this study, the ChatGPT (OpenAI, 2025) web interface and the GPT model Corpus Linguist (Milička & Machálek, 2024) were used. The Corpus Linguist model enables communication with the resources of the Czech National Corpus (CNC) project (Křen et al., 2016), which is an umbrella term for many corpora, not only those delivering data in the Czech language. The GPT model communicates through an API with written corpora (SYN2020, SYN v12), spoken corpora (ORTOFON), and parallel corpora (InterCorp) covering most European languages such as Croatian, Czech, Danish, Dutch, German, English, Finnish, French, Greek, Hungarian, Italian, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish, and Swedish. It also allows communication with the corpus of Czech contemporary poetry, the British National Corpus, and the Old Bailey Corpus. Of course, the CNC offers many more corpora available via its website (www.korpus.cz), but only the aforementioned ones are accessible through the API.

## 3 Workflow with Chat-GPT

In this study, the automatic generation of corpus-based exercises makes use of Czech corpora to evaluate exercises intended for foreign learners of Czech. However, all communication with the GPT model was carried out in English. The first step was to determine what kind of corpus-based exercises the model is able to generate. Next, the workflow with the model focused on a single exercise type in order to evaluate the interaction between the language model and corpus data.

Based on previous studies (Lin, 2025), it is known that the more detailed prompt is, the more plausible the results are. Therefore, communication with the Corpus Linguist model started with a detailed question: *Could you list the types of language exercises for foreigners that you are able to create based on corpus data?* In return, six categories of exercises were obtained:

1. Vocabulary & Collocations
   - Fill-in-the-blank with collocation

     example: "He made a ___ decision" (from frequent collocates of decision).
   - Match words with their typical partners

     example: *strong tea* vs. **powerful tea*.
   - Frequency-based vocabulary lists

     most common words in a domain, genre, or register.
2. Grammar Practice
   - Conjugation & declension in context sentences from corpus with a verb missing (*oni* ___ *domů* → *šli*, 'they went home').
   - Case exercises (Czech)

     fill the missing case ending from authentic examples.
   - Word order practice jumbled authentic sentences to be reordered correctly.
3. Reading & Comprehension
   - Gap-fill reading texts

     real corpus sentences with systematic gaps (e.g., prepositions, verbs).
   - Multiple-choice from concordances choose which option fits the authentic sentence.
   - Register awareness

     compare informal spoken excerpts (ORTOFON) with formal written ones (SYN).
4. Error Detection & Correction
   - Spot the odd one out

     one untypical collocate among three authentic ones.
   - Learner error simulations

     based on what corpus shows is not used, learners choose the correct form.
5. Translation & Contrastive Tasks (using InterCorp)
   - Parallel text exercises

     learners match Czech sentences with their English (or other language) equivalents.

- False friends & tricky equivalents

  highlight differences between corpora in two languages.

6. Communicative / Production

- Prompts with authentic phrases

  learners continue a sentence taken from corpus.

- Roleplay triggers

  situational expressions mined from spoken corpora ("Můžu si přisednout?", 'Can I sit down?').

The model indicated that all these exercises can be data-driven and offered to prepare a sample exercise. To ensure that all possible exercises were listed, the model was asked: *Did you listed all possible exercises?* In response, the model honestly answered that it was a representative set of exercises and that more possibilities are available depending on teaching goals, learners' level, and other factors. It then provided additional exercise types as follows:

1. Lexical / Phraseological

- Idioms & proverbs

  find authentic variants and ask learners to complete or explain them.

- Synonym choice

  pick the most frequent synonym in a given context.

2. Morphosyntax

- Transformation tasks

  rewrite corpus sentences (active → passive, direct → reported speech).

- Agreement practice

  highlight mismatches and ask learners to fix them (subject–verb, adjective–noun).

- Tense/aspect recognition (Czech perfective vs. imperfective verbs).

3. Discourse / Pragmatics

- Dialogue completion

  fill missing turns in authentic spoken dialogues.

- Register/style awareness

  classify sentences as formal/informal.

- Connector use

  identify or insert discourse markers (*totiž*, *nicméně* 'however').

4. Writing Support

- Sentence starters

  give frequent corpus-based openings, learners continue.

- Chunk recycling

  learners create new sentences with frequent multiword units.

- Error-avoidance training

  contrast common learner mistakes with authentic usage.

5. Assessment & Awareness

- Acceptability judgments

  choose the corpus-attested variant.

- Frequency awareness

  rank synonyms according to corpus frequency.

- Guess the meaning from context

  show concordances with a new word hidden.

Two exercise types were chosen for the further evaluation of the interaction between the language model and corpus data. The first was an exercise on error detection, where a learner needs to identify one untypical collocate among three authentic ones, and the second focused on false friends. The model was also asked to provide information about corpus data use.

Table 1 and Table 3 present a generative result of GenAI, divided into three steps: 1) *Exercise*, 2) *Task*, and 3) *Answer*. In both cases, a target word was not specified, so the model selected a lemma on its own.

In the first exercise, the chosen lexeme was *rozhodnutí* 'decision'. The GPT model explained the individual steps, stating that the top authentic collocates were provided based on the SYN2020 corpus and that one example was added as the "odd one out", an unattested collocate. However, a discrepancy can be observed between the collocates in the Exercise and Task parts. The collocation *učinit rozhodnutí* ('make a decision') does not appear in the Task part and is replaced by *těžké rozhodnutí* (‘hard decision’).

| | |
|---|---|
| **Prompt** | *Could you generate an exercise type "Spot the odd one out"? Please inform me, what kind of corpus data you used.* |
| **Exercise**: Spot the odd one out (Czech collocations) | |
| Target word: lemma *rozhodnutí* (“decision”).<br>• From the SYN2020 written corpus, I can extract its most frequent collocates.<br>• Top authentic collocates include: *učinit rozhodnutí* (“make a decision”), *důležité rozhodnutí* (“important decision”), *správné rozhodnutí* (“right decision”).<br>• I will then add one implausible or unattested collocate as the “odd one out”. | |
| **Task** | |
| Which of the following combinations is **NOT typical** in Czech? Circle the odd one out.<br>1. **důležité rozhodnutí**<br>2. **správné rozhodnutí**<br>3. **červené rozhodnutí**<br>4. **těžké rozhodnutí** | |
| **Answer** | |
| 3. **červené rozhodnutí** (“red decision”)<br>• It is not attested in corpus collocations and does not make semantic sense. | |

Table 1: Generative result of corpus exercise "Spot the odd one out".

To verify that the model was indeed communicating with the corpus, the most frequent collocates of the lexeme *rozhodnutí* ‘decision’ were extracted through the corpus portal in the same way as described by the model, using the logDice collocation measure, a span window of ±5 words, and a minimum frequency of 3 within the collocation. *Správné* (‘right’) was the only collocate provided by GenAI in the exercise that also appeared in the top 10 collocates (Table 2) based on the corpus search. It ranked 8th, while the next two appeared at ranks 43 (*důležité* ‘important’) and 62 (*těžké* ‘hard’).

The model stated that it “chose 3 real frequent collocates and added 1 fabricated/unattested one as the distractor.” However, not all collocates were among the most frequent. Therefore, the model was asked to deliver the collocation profile of the lemma *rozhodnutí* from the written Czech corpus SYN2020 (Křen et al., 2020). The results were identical to those of the corpus search, which confirmed the communication between the GPT model and the corpus portal.

| **Rank** | **Lemma** | **Traslation (POS)** | **Fq** | **logDice** |
|---|---|---|---|---|
| 1. | soud | court (N) | 940 | 9.72 |
| 2. | územní | territorial (A) | 305 | 9.11 |
| 3. | učinit | to make (V) | 354 | 9.08 |
| 4. | soudní | judicial (A) | 267 | 8.73 |
| 5. | padnout | to fall (V) | 303 | 8.66 |
| 6. | dospět | to come (V) | 174 | 8.18 |
| 7. | pravomocný | authoritative (A) | 135 | 8.17 |
| 8. | správný | right (A) | 285 | 8.08 |
| 9. | činit | to make (V) | 190 | 8.03 |
| 10. | konečný | final (A) | 203 | 8.01 |

Table 2: Top 10 collocate of lexeme *rozhodnutí* in SYN2020.

The most top collocates of lemma *rozhodnutí* are mostly related to an institutional context, which might be the reason why the model did not use this vocabulary, as it is not always pedagogically useful for lower proficiency levels. The model also argued that semantically clear collocations are needed for pedagogical reasons. This reasoning is valid, and the procedure for the automatic generation of corpus-based exercises provides not only authentic examples for learners but also the most appropriate ones from a pedagogical perspective. The longer the communication with the GPT model is conducted, the better the results obtained, and the interaction with corpus data is well explained.

In the second exercise (Table 3), the model chose the interesting Czech lexeme *eventuálně*, that is a false friend with English lexeme *eventually*. It explained that the parallel corpus InterCorp v13ud (Čermák & Rosen, 2012) was used to extract concordance lines for *eventuálně* in Czech texts aligned with English translations. The model also assumed that corpus frequency results identified *possibly* and *alternatively* as the major equivalents. However, a direct consultation of the corpus revealed discrepancies. Following the same steps described by the model, 156 hits of *eventuálně* were found in the Czech-English part of InterCorp v13ud (Klégr et al., 2021). Manual analysis of the context confirmed that the most frequent equivalent was *possibly*, with 33 hits, however, it was followed by *eventually* with 28 (in the subtitle collection 14) hits (Figure 1). *Alternatively* appeared only 7 times. The other results were mostly translated by omission or by other equivalents not mentioned by the model. Further prompting revealed that the model is not able to consult corpus data when it comes to two parallel corpora aligned directly. The current API settings do not allow this function. In fact, the model provided misleading information, claiming that the corpus data delivered the most frequent translations. This type of hallucination is particularly problematic, as the Corpus Linguist model was expected to cooperate with corpus data. Nonetheless, it highlights the unreliability of some outputs, which users should be aware of.

<table>
<tr><td><b>Prompt</b></td><td><i>Could you generate an exercise type "False friends & tricky equivalents"? Please inform me, what kind of corpus data you used.</i></td></tr>
<tr><td colspan="2"><b>Exercise</b>: Exercise: False Friends (Czech–English)</td></tr>
<tr><td colspan="2">Target word: Czech <i>eventuálně</i><br>• Many learners think it means English <i>eventually</i>.<br>• But in corpus data, <i>eventuálně</i> is most often translated as <i>possibly / alternatively</i>.</td></tr>
<tr><td colspan="2"><b>Task</b></td></tr>
<tr><td colspan="2">Fill in the blanks with the correct English equivalent of Czech <i>eventuálně</i>. Choose between <b>eventually</b> and <b>possibly</b>.<br><i>Pokud se nám to nepodaří, můžeme ___ zvolit jinou metodu.</i><br>→ If we don’t succeed, we can ___ choose another method.<br><i>Po dlouhých diskusích ___ dospěli k rozhodnutí.</i><br>→ After long discussions, they ___ reached a decision.<br><i>Tyto části mohou být ___ nahrazeny novými.</i><br>→ These parts can be ___ replaced by new ones.</td></tr>
<tr><td colspan="2"></td></tr>
<tr><td colspan="2"><b>Answer</b></td></tr>
<tr><td colspan="2">1. possibly<br>2. eventually<br>3. possibly</td></tr>
</table>

Table 3: Generative result of corpus exercise "False friends & tricky equivalents".

The exercise itself asks learners to fill in the gaps in English sentences, not in Czech ones. Therefore, it resembles practice for Czech native speakers studying English rather than exercises for foreign learners of Czech. Moreover, the second Czech sentence should not contain the lexeme *eventuálně* but *nakonec* ‘in the end’. This exercise requires improvement in this regard. Further prompting produced a new version of the exercise in which learners were asked to choose between two English sentences, e.g. **The trip can eventually be extended* and *The trip can possibly be extended*, to identify the correct equivalent of the given Czech sentence *Cestu lze eventuálně prodloužit*.

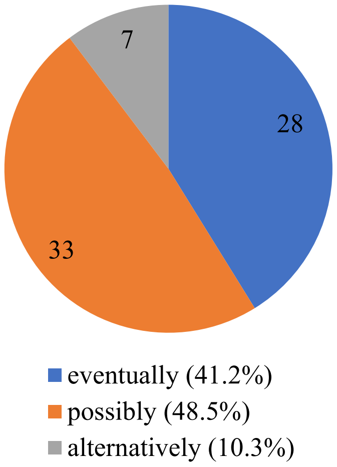

Figure 1: Distribution of translation equivalents of lexeme *eventuálně* in InterCorp 13ud.

## 4 Discussion

This study, based on two examples of corpus-based exercises, explored the possible communication of a GPT model (Corpus Linguist) with corpus data (CNC). The results revealed that the model successfully extracted data from corpora in some cases, however, in other cases, it claimed to use corpus data that were not actually consulted. Comparison with previous a study (Zasina, 2025) also indicates that the Corpus Linguists model tends to choose similar target lexemes (*rozhodnutí* ‘decision’) for different exercises when a target lexeme is not specified in the prompt. The model appears to be pre-set to choose target words for exercise generation, which represents a certain limitation of the tool. Therefore, for the automatic generation of corpus-based exercises, it is essential to deliver input data concerning the most problematic areas for the learners for whom the exercises are intended. This precaution has the potential to improve the generated results and better target learner-specific needs. It seems that the model is not yet able to truly consider learners’ real needs, and this remains the task of materials developers.

The two examples also demonstrated that it is crucial to define detailed prompts and verify the answers through subsequent prompting. Longer interactions make it possible to identify the processes undertaken by the GPT model. This is an important consideration in prompt crafting, to avoid succumbing to the illusion that every answer is correct. It should be remembered that GenAI can only produce strings of characters that form words and sentences; it is not capable of independent thought.

An interesting concept of GenAI as a role player (see Shanahan et al., 2023) can be applied in this context. When prompting a GPT model, one may feel that the automatically generated sentences are human-like. However, GenAI's ability to "act" convincingly stems from its vast training data (Shanahan et al., 2023, p. 496), and it strives to deliver coherent responses. Some scholars (Milička, 2024, p. 16) have emphasised that the model should not be anthropomorphised. Anthropomorphising GPT models can be a trap that dulls vigilance. Thus, it is important to critically assess GenAI's outputs, especially when combining them with corpus data.

Some attempts to integrate GenAI with corpus data in language learning have been undertaken (Cheung & Crosthwaite, 2025; Crosthwaite & Anthony, 2025, p. 6; Zasina, 2025) and have produced promising results. However, this study highlighted its limitations regarding communication with corpus data via the ChatGPT interface, which tends to provide an answer under any circumstances. Users should therefore be cautious and prepare more elaborated prompts that explicitly query the source data.

Even though this evaluation is limited to two examples of corpus-based exercises, it provides insight into how GenAI interacts with CNC sources. It offers guidance for future developments in prompt crafting and for evaluating the reliability of generative results. Furthermore, it underscores that users should approach GenAI critically. Further investigation may lead to improvements that could mitigate these issues.

## 5 Conclusion

This study focused on the interaction between language models and corpus data in terms of DDL practice. It shows that there are many possibilities for language learners to combine GenAI and corpus data. GenAI can effectively use corpus data within GPT models to produce corpus-based exercises. However, it is important to emphasise that prompts should contain high-quality input information and be as precise as possible in order to obtain appropriate results. It is also crucial to critically evaluate the automatically generated outputs, as they can be hallucinated by GenAI. In the future, further evaluations of this kind will be necessary to determine whether AI sufficiently cooperates with corpus data or merely creates the illusion of real corpus results.

## Acknowledgments

This work was supported by the project "Human-centred AI for a Sustainable and Adaptive Society" (reg. no.: CZ.02.01.01/00/23_025/0008691), co-funded by the European Union, and the project "Multilingual Lens: Investigating Large Text Corpora from Different Methodological Perspectives" (UNCE/24/SSH/009).

## References

Čermák, F., & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, *17*(3), 411–427.

Cheung, L., & Crosthwaite, P. (2025). *CorpusChat*: Integrating corpus linguistics and generative AI for academic writing development. *Computer Assisted Language Learning*, 1–27. https://doi.org/10.1080/09588221.2025.2506480

Crosthwaite, P., & Anthony, L. (2025). Tools for Data-Driven Learning. In *The Palgrave Encyclopedia of Computer-Assisted Language Learning* (pp. 1–9). Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-031-51447-0_74-1

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. *Classroom Concordancing: ELR Journal*, *4*, 1–16.

Klégr, A., Kubánek, M., Malá, M., Rohrauer, L., Šaldová, P., Šebestová, D., Vavřín, M. & Rosen A. (2021). *InterCorp – English, Release 13ud of 22 December 2021*. Institute of the Czech National Corpus, Charles University. www.korpus.cz

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kováříková, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondřička, P. & Zasina, A. J. (2016). SYN2015: Representative Corpus of Contemporary Written Czech. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 2522–2528). Portorož: ELRA.

Křen, M., Cvrček, V., Henyš, J., Hnátková, M., Jelínek, T., Kocek, J., Kováříková, D., Křivan, J., Milička, J., Petkevič, V., Procházka, P., Skoumalová, H., Šindlerová, J., & Škrabal, M. (2020). *SYN2020: Representative corpus of written Czech*. Institute of the Czech National Corpus, Faculty of Arts, Charles University. www.korpus.cz

Lin, P. (2025). AI Chatbots and Data-Driven Learning. In L. McCallum & D. Tafazoli (Eds.), *Encyclopedia of Computer-Assisted Language Learning* (pp. 1–8). Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-031-51447-0_67-1

Milička, J. (2024). *Theoretical and Methodological Framework for Studying Texts Produced by Large Language Models* (arXiv:2408.16740). arXiv. https://doi.org/10.48550/arXiv.2408.16740

Milička, J., & Machálek, T. (2024). *Corpus Linguist* (Version January 17, 2025) [Computer software]. https://chatgpt.com/g/g-pFqRCNeHu-corpus-linguist

OpenAI. (2025). *Chat-GPT* (Version September 9, 2025) [Computer software]. https://chatgpt.com/

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, *623*(7987), 493–498. https://doi.org/10.1038/s41586-023-06647-8

Vyatkina, N. (2020). Corpora as open educational resources for language teaching. *Foreign Language Annals*, *53*(2), 359–370. https://doi.org/10.1111/flan.12464

Zasina, A. J. (2022). Designing a Corpus Workbook for Students of Czech as a Foreign Language. *Studie z Aplikované Lingvistiky - Studies in Applied Linguistics*, *13*(2), 125–132.

Zasina, A. J. (2025). Typologie korpusových cvičení a jejich automatické generování pomocí AI [Typology of corpus-based exercises and their automatic generation using AI]. In M. Škrabal, B. Štěpánková & H. Skoumalová (Eds.), *Korpus třicetiletý* (pp. 127−145). Praha: Nakladatelství Lidové noviny.